Analysis of intervariable relationships between major risk factors in the development of coronary artery disease: a classification tree approach

Koroner arter hastalığı gelişiminde majör risk faktörlerinin birbirleri arasındaki ilişkilerin incelenmesi: Bir sınıflandırma ağacı yaklaşımı

*Mevlüt Türe, İmran Kurt§, Turhan Kürüm**

From Departments of Biostatistics and *Cardiology, Medical Faculty, Trakya University, Edirne
Department of §Biostatistics, Medical Faculty, Eskişehir Osmangazi University, Eskişehir, Turkey

ABSTRACT

Objective: The purpose of this study is to determine how the major risk factors are related to each other in the development of coronary artery disease (CAD) using Chi-squared Automatic Interaction Detection (CHAID).

Methods: All patients with suspected CAD seen in the cardiology clinic between January 1999 and February 2003 who underwent coronary angiography were included in the study. A retrospective analysis was performed in 1381 patients. In all patients' sex, age, type II diabetes mellitus, hypercholesterolemia, systemic hypertension, smoking status, family history of CAD, body mass index (BMI) were assessed.

Results: According to classification tree, first-level split produced the two initial branches: female (unadjusted presence percentage = 48.07%) versus male (unadjusted presence percentage = 78.02%). For the male aged between 49-81 years and the female aged between 15-48, 49-60 and 61-71 years, diabetes mellitus was the most prominent risk factor. However, hypercholesterolemia was the best predicting variable for the females aged between 72-81 years. For the females of 15-48 years and 49-60 years age categories without diabetes mellitus, smoking status and family history of CAD had important contribution to the model.

Conclusion: Sorting the major risk factors of CAD from the most to least according to the classification importance was resulted as sex, age, diabetes mellitus, hypercholesterolemia, family history of CAD and smoking status. *(Anadolu Kardiyol Derg 2007; 7: 140-5)*

Key words: CHAID, coronary artery disease, risk factors, classification tree

ÖZET

Amaç: Bu çalışmanın amacı, koroner arter hastalığının (KAH) gelişmesine etki ettiği bilinen majör risk faktörlerinin bir birleri ile ilişkilerinin nasıl olduğunu "Chi-squared Automatic Interaction Detection" (CHAID) metodu kullanarak belirlemektir.

Yöntemler: Ocak 1999-Şubat 2003 yılları arasında kardiyoloji kliniğinde KAH'dan şüphelenilen ve koroner anjiyografisi yapılan bütün hastalar çalışmaya alınmıştır. Geriye dönük araştırma 1381 hasta üzerinde yapılmıştır. Bütün hastalar için cinsiyet, yaş, tip II diyabet, hiperkolesterolemi, sistemik hipertansiyon, sigara kullanımı, ailede KAH olma, vücut kitle indeksi kullanıldı.

Bulgular: Sınıflandırma ağacında, birinci bölünmeyle iki başlangıç dal meydana geldi: kadınlarla erkeklerin düzeltilmemiş KAH olma yüzdesi sırasıyla %48.07 ve %78.02 bulundu. Yaşı 49-81 arasında olan erkekler ve yaşı 15-48, 49-60 ve 61-71 arasında olan kadınlarda diyabet en önemli faktör olarak bulundu. Bununla birlikte, hiperkolesterolemi, yaşı 72-81 arasında olan kadınlar için en iyi kestirici olarak bulundu. Yaşı 15-48 ve 49-60 arasında olan diyabetsiz kadınlar için sigara kullanımı ve ailede KAH olma kestirici değişkenlerinin modele önemli katkıları olduğu görüldü.

Sonuç: Bu çalışmada, KAH riski üzerine etkisi olan faktörler sınıflandırma önemine göre cinsiyet, yaş, diyabet, hiperkolesterolemi, ailesinde KAH olma ve sigara kullanımı şeklinde sıralanmıştır. *(Anadolu Kardiyol Derg 2007; 7: 140-5)*

Anahtar kelimeler: CHAID, koroner arter hastalığı, risk faktörleri, sınıflandırma ağacı

Introduction

Coronary artery disease (CAD) is a major worldwide health problem with its incidence and mortality rates (1).

Many risk factors are known to play role in pathogenesis of CAD and myocardial infarction. Family history, smoking, hypertension, hypercholesterolemia, and diabetes mellitus have been described as the major risk factors for CAD (2-6). Identification of risk factors in CAD is essential for the management and follow-up of CAD patients. Numerous cross-sectional angiography studies have reported correlations of one or more major risk factors for myocardial infarction with the presence of CAD. How well these risk factors correlate with the degree of CAD is not clear.

Address for Correspondence: Mevlut Türe, Assistant Professor, Trakya University Faculty of Medicine Department of Biostatistics 22030 Edirne, Turkey
Tel.: +90 284 235 76 41/1631 Fax: +90 284 235 76 52 E-mail: ture@trakya.edu.tr

The aim of this study is to present an induction technique as a data mining approach to determine how major known risk factors related to each other in the development of CAD using induction technique.

Methods

Patients

All patients with suspected CAD seen in the cardiology clinic between January 1999 and February 2003 who had coronary angiography because of stable angina pectoris or an acute coronary syndrome or atypical angina were included in the study. A retrospective analysis was performed in 1381 patients (992 male (71.83%), 389 female (28.17%)) with suspected CAD. Clinically relevant CAD was defined by the presence of at least one vessel with a stenosis ≥50% in coronary angiography to be done due to angina associated with evidence for myocardial ischemia either by stress electrocardiography, stress Tc99 MIBI scintigraphy or pathologic resting electrocardiography. Patients with non-atherosclerotic CAD were not included into the study. In all patients' sex, age and all major risk factors including type II diabetes mellitus, hypercholesterolemia, systemic hypertension, smoking status, family history of CAD, body mass index (BMI) were assessed and documented. All angiograms were assessed by 2 cardiologists not participating in the study.

Data

Hypertension was diagnosed when the systolic blood pressure (BP) was≥140 mm Hg and/or diastolic BP was ≥90 mm Hg on at least three separate occasions and was established by the absence of clinical findings suggestive of secondary form of hypertension (2). Blood pressure was measured in the sitting position in a quite room, using a mercury sphygmomanometer, after the patient had rested for at least 10 min. Systolic BP was recorded at the appearance of sounds (Korotkoff phase I) and the diastolic at their disappearance (Korotkoff phase V). Patients who currently smoked or discontinued smoking during the last 6 months were categorized as smokers. Diabetes was confirmed if the questionnaire indicated one of the following National Diabetes Data Group criteria: 1- Symptoms of diabetes plus casual plasma glucose concentration 200 mg/dl (11.1 mmol/l). Casual is defined as any time of day without regard to the time since last meal. The classic symptoms of diabetes include polyuria, polydipsia, and unexplained weight loss. 2- Fasting plasma glucose (FPG) level ≥126 mg/dl (7.0 mmol/l). Fasting is defined as no caloric intake for at least 8 hours. 3- 2-h PG ≥200 mg/dl (11.1 mmol/l) during an oral glucose tolerance test (OGTT). The test should be performed as described by WHO (2), using a glucose load containing the equivalent of 75 g anhydrous glucose dissolved in water (3). Hypercholesterolemia was accepted when the total cholesterol was above 200 mg/dl (4). Measurement of routine laboratory examinations of serum and urine were performed in the fasting state.

We analyzed the simultaneous relationship among risk factors for CAD. This study compared the relative effects of each risk factor for CAD in the multivariate analysis model. We tried to discover the significant patterns and relationship among the risk factors and make decision rules for the management of CAD.

Comparison of characteristics

The characteristics of the study population are shown in Table 1. We performed the classical statistical analysis to examine the difference in the distribution of variables between the presence and absence of CAD. Age and BMI were tested for normal distribution by the Kolmogorov-Smirnov test. Comparison between any two groups was performed by unpaired t test for normally distributed data or non-parametric Mann-Whitney U test for non-normally distributed data. Nominal variables were tested by Chi-square test for presence and absence of CAD.

Five variables (sex, age, smoking status, diabetes mellitus and family history of CAD) were independently significant between presence and absence of CAD (p<0.001, p<0.001, p<0.001, p<0.001 and p=0.026, respectively).

Classification tree

A classification tree is a non-linear discrimination method, which uses a set of independent variables to split a sample into progressively smaller subgroups. The procedure is iterative at each branch in the tree, it selects the independent variable that has the strongest association with the dependent variable according to a specific criterion (7-9). The classification tree assumes that the effect of a variable in the subset is unrelated to the effect of the variable in other subsets of subjects.

The Chi-squared Automatic Interaction Detection (CHAID), a method that uses Chi-squared statistics to identify optimal splits, was used in this study. A CHAID tree is a classification tree that is constructed by repeatedly splitting subsets of the space into two or more child nodes, beginning with the entire data set (8,9). To determine the best split at any node, any available pair of categories of the predictor variables is merged until there is no statistically significant difference within the pair with respect to the target variable. This CHAID method naturally deals with interactions between the independent variables that are directly available from an examination of the tree. The final nodes identify subgroups defined by different sets of independent variables.

Cross-validation involves splitting the sample into a number of smaller samples. Trees are then generated, excluding the

Table 1. Characteristics of study subjects

Characteristics	CAD		p
	Presence (n=961)	Absence (n=420)	
Sex, male/female	4.1	1.2	<0.001
Age, years	60 (51-67)	55 (46-64)	<0.001
BMI, kg/m^2	26.8±3.8	27.3±4.6	NS
Smoking status, %	59.6	43.4	<0.001
Diabetes mellitus, %	22.5	10.5	<0.001
Systemic hypertension, %	45.9	41.6	NS
Hypercholesterolemia, %	46.3	42.8	NS
Family history of CAD, %	24.8	19.6	0.026

Continuous variables are mean±SD; Smoking status: current or discontinued during the last 6 months; Diabetes mellitus: HbA1c ≥6.2%, current insulin, or oral hypoglycemic therapy; Systemic hypertension: >140/90 mmHg or current antihypertensive therapy; Hypercholesterolemia: cholesterol ≥200 mg/dl or current lipid-lowering drugs;

BMI- body mass index, CAD- coronary artery disease, NS- nonsignificant data from each subsample in turn. For each tree, misclassification risk is estimated by applying the tree to the subsample excluded in generating it. The cross-validated risk estimate for the overall tree is calculated as the average of the risks for all of these trees (10).

In this paper, the CHAID algorithm with growing criteria of the likelihood ratio Chi-square statistic was used for building the tree and evaluating the splits. To identify the nodes with a relatively high probability, a gain chart was constructed showing the nodes sorted by the number of cases in the target category for each node.

All statistical analyses were performed using SPSS and ANSWER TREE statistical software (SPSS Inc. Chicago, Il, USA).

Results

Classification tree and rules for the prediction of CAD

Of the eight variables that were entered in CHAID analysis, six were selected by the program for the classification tree. The six variables were: sex, age, diabetes mellitus, hypercholesterolemia, smoking status and family history of CAD.

The CHAID identified the variables that play important role in explaining presence of CAD (Fig. 1). This indicated that the sex was the most important determining factor. This first-level split produced the two initial branches of the classification tree: female (unadjusted presence percentage = 48.07%) versus male (unadjusted presence percentage = 78.02%).

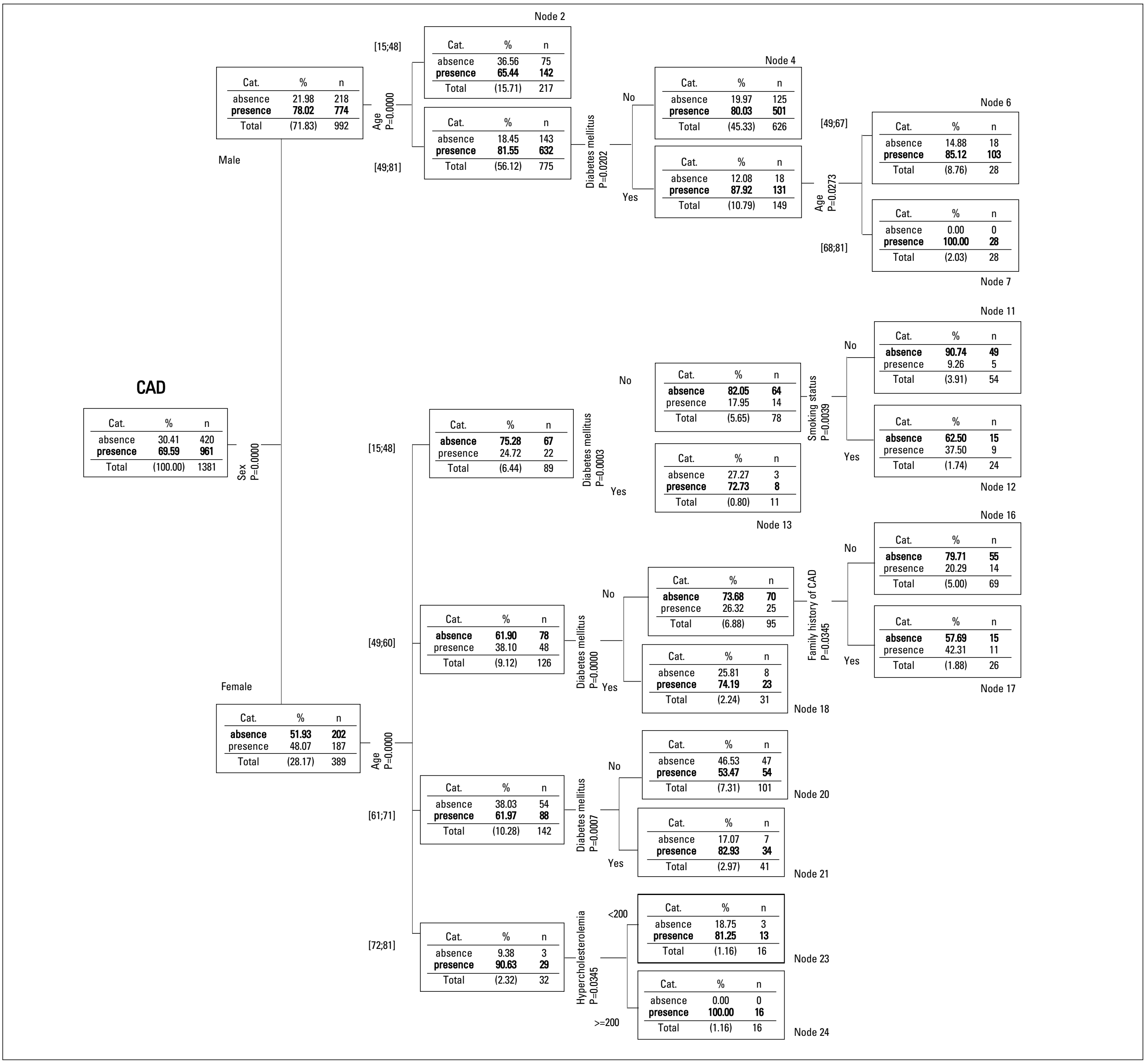

Figure 1. Decision tree by CHAID algorithm

CAD- coronary artery disease

We could see differences in two subtrees. For the both female and male, age proved to be the best predicting variable. For the males aged between 49-81 years, diabetes mellitus was the most prominent. Also, for the female and three difference categories of age (15- 48, 49-60 and 61-71 years), diabetes mellitus was the most prominent variable. However, hypercholesterolemia was the best predicting variable for the females aged between 72-81 years. For the males aged between 49-81 years with diabetes mellitus, the age was the significant variable. For the females of 15-48 years and 49-60 years age categories without diabetes mellitus, smoking status and family history of CAD also added useful information to the model.

Classification trees are charts that illustrate decision rules (Table 2). The decision rules provide specific information about risk factors based on the rule induction. They begin with one root node that contains all of the observations in the sample. As shown in Figure 1, the classification tree has 24 leaf nodes, of which 14 are terminal nodes. Each node depicted in the classification tree can be expressed in terms of an 'if-then' rule.

Target segmentation for the management of CAD

The gains chart produced by the classification tree can be used for a risk analysis for presence of CAD management. The gain summary shows, which nodes have the highest and the lowest proportions of a target category within the node. There are two parts of the gains chart: node-by-node statistics and cumulative statistics (Table 3). In the gains chart, nodes were sorted by the number of cases in the target category for each node.

The first node in the Table 3, node 7, contains 28 presence of CAD out of 28 subjects, i.e. a presence rate of 100.00%. For this type of gains chart, with a categorical target variable, the gain score equals the percentage of cases with the target category-in this case, presence of CAD-for the node. The index score shows how the proportion of presence for this particular node compares to the overall proportion of presence of CAD. For node 24, the index score is about 143.70%, indicating that the proportion of respondents for this node is about 1.4 times the presence of CAD rate for the overall sample. The cumulative statistics demonstrate how well we do at finding presence cases by taking the best segments of the sample. If we only take the best node (node 7), we reach 2.91% of presence cases by targeting only 2.03% of the sample. If we include the next best node as well (node 24), then we get 4.58% of the presence from only 3.19% of the sample. Including node 6 increases these values to 15.30% of presence cases from 11.95% of the sample. If we include until node 13, we get 75.55% of presence cases and we must contact 64.45% of the sample to get them. At this stage, we are at the crossover point described above, where we start to see diminishing returns.

The gains chart also provides valuable information, which segments to target, and which to avoid. We might base the decision on the number of prospects we want, the desired presence rate for the target sample, or the desired proportion of all potential presence cases we want to contact.

The overall risk estimate in classification tree was 0.2353 (standard error of risk estimate 0.0114), indicating that 76.47% of the cases will be classified correctly by using the decision rule based on the current tree. However, the cross-validated risk estimate was 0.2411 (standard error of risk estimate 0.0115). Sensitivity, specificity and predictive rate calculated by CHAID model were found to be 95.9%, 31.9%, and 76.5% respectively.

Discussion

Cardiovascular disease is the leading cause of mortality in Turkey as in the world (11). The most important recent epidemiological investigation in Turkish population, the Turkish Adult Risk Factors Study, has shown that the incidence of CAD in young individuals is common in this population (12).

Cardiovascular disease is generally due to combination of several risk factors. Risk factors modifications have been unequivocally shown to reduce mortality and morbidity, especially in people with either unrecognized or recognized cardiovascular disease (1).

Table 2. Decision rules for the prediction of presence of CAD

Node	Sex	Age, years	Diabetes mellitus	Hypercholesterolemia	Smoking status	Family history of CAD	Probability of presence, %
7	male	67<age≤81	Yes	*	*	*	100.00
24	female	71<age≤81	*	≥200	*	*	100.00
6	male	48<age ≤67	Yes	*	*	*	85.12
21	female	60<age≤71	Yes	*	*	*	82.93
23	female	71<age≤81	*	<200	*	*	81.25
4	male	48<age≤81	No	*	*	*	80.03
18	female	48<age≤60	Yes	*	*	*	74.19
13	female	15≤age≤48	Yes	*	*	*	72.73
2	male	15≤age≤48	*	*	*	*	65.44
20	female	60<age≤71	No	*	*	*	53.47
17	female	48<age≤60	No	*	*	yes	42.31
12	female	15≤age≤48	No	*	yes	*	37.50
16	female	48<age≤60	No	*	*	no	20.29
11	female	15≤age≤48	No	*	no	*	9.26

*- Nonsignificant
CAD- coronary artery disease

Guo et al. (13) evaluated the correlation between multiple cardiovascular risk factors and the extent and severity of angiographic coronary artery disease in patients underwent coronary angiography. They reported that following cardiovascular risk factors were included: age, gender, hypertension, smoking status, type-2 diabetes mellitus, dyslipidemia, and high uric acid level. We used classification tree by CHAID algorithm, and examined the relationship among demo-graphic and clinical factors for CAD and found the different priority among conventional risk factors.

We also showed how CHAID could be used in risk analysis and target segmentation for CAD management. Using the CHAID algorithm, we obtained cumulative statistics, which show how well we found the presence cases by taking the best segments of the sample. The gains chart also provided valuable information, which segments to target and which to avoid. Besides, we presented the rules that provided an occurrence relationship among the risk factors. This type of information can be used in analyzing the individual risk factors' effects on a specific segment of the target population. While hypercholesterolemia was not found to be the important risk factor of CAD using univariate test, it was found to be the important determining factor by CHAID method.

The presence of any risk factor alone or in combination with others (any of two or all three of smoking, serum total cholesterol, and systolic blood pressure) in the Multiple Risk Factor Intervention Trial was associated with the steeper increase in cardiovascular disease mortality for diabetic and non-diabetic patients (14).

Patients with diabetes are at higher mortality risks than those without diabetes (15). Kempler concluded that management of cardiovascular risk factors should be considered as the cornerstone of diabetes care (16).

Cigarette smoking is a major risk factor for total mortality in the general population (17). Mortality rates among women with diabetes are strongly related to their smoking habits, with heavier smokers having a twofold higher risk than non smokers (18).

In the decision tree (Fig. 1), sex was the most important determining factor on first-level split. Age was the most important determining factor on the second-level split. These factors were followed by diabetes mellitus and hypercholesterolemia showing the third order of importance. Furthermore smoking and family history of CAD were found to have lesser importance than these risk factors.

Kim et al. (19) found that age and diabetes mellitus significantly affected CAD for male and female but BMI, family history of CAD and hypertension were not significant predictors in their study of 544 subjects. Furthermore, they found that smoking was not significant risk factor for male but it was significant factor for female. Rahman et al. (5) reported in 112 subjects that sex, smoking and family history of CAD and hypertension were important risk factors of CAD. In our study, smoking status was more important in 15-48 age groups for females, and family history of CAD was more important factor for 48-60 age groups for females. Interestingly, BMI was excluded by CHAID.

We found a high sensitivity (95.9%) but a lesser specificity (31.9%). This may be due to several limitations of the study. Since it is relied on data from a single hospital, the study was biased in terms of the affluence of the subject group. The other limitations were the lack of input variables for risk factors, such as exercise behavior, lipoprotein (a), hyperuricemia and homocysteinemia.

In conclusion, this study indicated that sex, age, diabetes mellitus, hypercholesterolemia, family history of CAD and smoking status are important CAD risk factors, and sex and increasing age are major, non-modifiable CAD risk factors.

References

1. Backer GD, Ambrosioni E, Borch-Johnsen K, Brotons C, Cifkova R, Dallongeville J, et al. European guidelines on cardiovascular disease prevention in clinical practice. Third Joint Task Force of European and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (constituted by representatives of eight societies and by invited experts). Eur Heart J 2003; 24: 1601-10.

Table 3. Gain chart by CHAID algorithm

Node	Node-by-Node						Cumulative					
	Node, n	Node, %	Resp., n	Resp., %	Gain, %	Index, %	Node, n	Node, %	Resp., n	Resp., %	Gain, %	Index, %
7	28	2.03	28	2.91	100.00	143.70	28	2.03	28	2.91	100.00	143.70
24	16	1.16	16	1.66	100.00	143.70	44	3.19	44	4.58	100.00	143.70
6	121	8.76	103	10.72	85.12	122.33	165	11.95	147	15.30	89.09	128.03
21	41	2.97	34	3.54	82.93	119.17	206	14.92	181	18.83	87.86	126.26
23	16	1.16	13	1.35	81.25	116.76	222	16.08	194	20.19	87.39	125.58
4	626	45.33	501	52.13	80.03	115.01	848	61.40	695	72.32	81.96	117.78
18	31	2.24	23	2.39	74.19	106.62	879	63.65	718	74.71	81.68	117.38
13	11	0.80	8	0.83	72.73	104.51	890	64.45	726	75.55	81.57	117.22
2	217	15.71	142	14.78	65.44	94.04	1107	80.16	868	90.32	78.41	112.68
20	101	7.31	54	5.62	53.47	76.83	1208	87.47	922	95.94	76.32	109.68
17	26	1.88	11	1.14	42.31	60.80	1234	89.36	933	97.09	75.61	108.65
12	24	1.74	9	0.94	37.50	53.89	1258	91.09	942	98.02	74.88	107.61
16	69	5.00	14	1.46	20.29	29.16	1327	96.09	956	99.48	72.04	103.53
11	54	3.91	5	0.52	9.26	13.31	1381	100.00	961	100.00	69.59	100.00

Resp.- Respondent

2. Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL Jr, et al. National Heart, Lung and Blood Institute Joint National Committee on Prevention, detection, Evaluation, and Treatment of High Blood Pressure; National High Blood Pressure Education Program Coordinating Committee. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood pressure: the JNC 7 report. JAMA 2003; 289: 2560-72.
3. The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care 2003; 26 (Suppl 1): S5-S20.
4. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). JAMA 2001; 285: 2486-97.
5. Rahman MA, Chaudhury HS, Malek MA, Khaled MA. Risk factors of angiographically defined coronary artery disease in Bangladesh. Cardiovasc Pathol 2004; 13 (Suppl): S80-S138.
6. Herrington DM, Kesler K, Reiber JHC, Davis W, Brown WV, Helms R, et al. Arterial compliance adds to conventional risk factors for prediction of angiographic coronary artery disease. Am Heart J 2003; 146: 662–7.
7. Ho SH, Jee SH, Lee JE, Park JS. Analysis on risk factors for cervical cancer using induction technique. Expert Syst Appl 2004; 27: 97-105.
8. Michael JAB, Gordon L. Data mining technique: for marketing, sales and customer support. 1st edition. New York: Wiley; 1997.
9. Ture M, Kurt I, Kurum AT, Ozdamar K. Comparing classification techniques for predicting essential hypertension. Expert Systems with Applications 2005; 29: 583-8.
10. Magidson J, SPSS Inc. SPSS for Windows CHAID Release 6.0. 2nd ed. Chicago: SPSS Inc.; 1993.
11. Tokgozoglu L, Pehlivanoglu S, Kultursay H, Oguz A, Damci T, Senocak M, et al. Which patients have the highest cardiovascular risk? A follow-up study from Turkey. Eur J Cardiovasc Prev Rehabil 2005; 12: 250-6.
12. Berdeli A, Sekuri C, Cam FS, Ercan E, Sagcan A, Tengiz I, et al. Association between the eNOS (Glu298Asp) and the RAS genes polymorphisms and premature coronary artery disease in a Turkish population. Clinica Chimica Acta 2005; 351: 87-94.
13. Guo YH, Zhang WJ, Zhou YJ, Zhao D, Zhou ZM, Zhang H. Study of the relationship between cardiovascular risk factors and severity of coronary artery disease in patients underwent coronary angiography. Zhonghua Xin Xue Guan Bing Za Zhi 2005; 33: 415-8.
14. Stamler J, Vaccaro O, Neaton JD, Wentworth D. For the Multiple Risk Factors Interventional Trial research group. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. Diabetes Care 1993; 16: 434-44.
15. Dierkx R, Van De Hoek W, Hoekstra J, Erkelens D. Smoking and diabetes mellitus. Neth J Med 1996; 48: 150-62.
16. Kempler P. Learning from large cardiovascular clinical trials: classical cardiovascular risk factors. Diab Res Clin Pract 2005; 68 (Suppl 1): S43-7.
17. Kawachi I, Colditz G, Stampfer MJ, Willett WC, Manson JE, Rosner B, et al. Smoking cessation in relation to total mortality rates in women. Intern Med 1993; 119: 992-1000.
18. Al-Delaimy WK, Willett WC, Manson JE, Speizer FE, Hu FB. Smoking and mortality among women with type-2 diabetes. The Nurses' Health Study Chort. Diabetes Care 2001; 24: 2043-8.
19. Kim HK, Chang SA, Choi EK, Kim YJ, Kim HS, Sohn DW, et al. Association between plasma lipids, and apolipoproteins and coronary artery disease: a cross-sectional study in a low-risk Korean population. Int J Cardiol 2005; 101: 435-40.